Bayesian Statistical Reanalysis of Red Dye 3 Thyroid Neoplasms

Lyle D. Burgoon, Ph.D., Fellow ATS

February 8, 2025

Bottom-Line Up-Front

I find that, and it is my opinion that:

- US FDA erred in granting the petitioners' remedy.
- US FDA improperly summed adenomas and carcinomas in direct violation of the Delaney Clause. My reasoning is based on the strict interpretation that the petitioners reminded US FDA they had to use under the 9th Circruit decision, and based on the more recent Loper Bright decision from the Supreme Court. The Delaney Clause is clear that US FDA only consider if a chemical causes "cancer". The Delaney Clause did not say that US FDA was to restrict chemicals that cause "benign tumors" or "adenomas". At the time the Delaney Clause was passed, there was a clear understanding of the differences between adenomas and carcinomas. As the 9th Circuit found that Congress was clear and unambiguous in its wording, it is clear that Congress meant cancers when it said "cancers" in the Delaney Clause.
- US FDA did not consider all of the data together – it still only considers the data separately. Thus, the US FDA is not considering the fact of regression towards the middle nor is US FDA considering the impacts of small sample sizes in creating false positive results. A reasonable scientist considers the total weight of the evidence, not any one, singular study on its own. Consideration of the weight of evidence is a standard practice in food and chemical risk assessment.
- US FDA should have considered the data that petitioners relied upon to be unreliable. The sample sizes are too small to be reliable predictors of the population response. They are too small to invoke the Central Limit Theorem. Thus US FDA erred in not declaring the data to be unreliable – which they have done in other Delaney Clause cases.

- The data are clear based on my analysis – the 4% group did not actually see an increase in carcinomas once you consider all of the data together, and once you consider the historical background rate of thyroid carcinomas in the vehicle animals.

Background

There is an ongoing controversy regarding whether red dye 3 (A.K.A. FD&C Red No. 3 and erythrosine) is a known animal carcinogen. Specifically, the Center for Science in the Public Interest et al. (the petitioners) filed a citizen petition with the US FDA alleging that FDA had already decided that FD&C Red No. 3 causes cancer in rats (CAP 3C0323; Federal Register 88 FR 10248). Therefore, the petitioners argued that the US FDA is compelled by law (the Delaney Clause) to repeal the color additive regulations for FD&C Red 3 in Section 74.303 (21 CFR 74.303) and Section 74.1303 (21 CFR 74.1303).

The argument from petitioners is focused on the fact that petitioners claim, and the US FDA seemingly agreed, that FD&C Red No. 3 causes a statistically significant increase in cancers of the thyroid in male Sprague Dawley rats.

The petitioners and US FDA relied upon data submitted to the US FDA, especially the data found in a US FDA memorandum dated 11 AUGUST 1989 from the Deputy Director of the Division of Toxicological Review and Evaluation to Ronald Lorentzen (memo 9C0096; the memo). This memo includes "data that may influence the CAncer Assessment Committee's decision whether FD&C Red No. 3 mediates thyroid tumorigenesis in the rat via a primary (direct) or a secondary (indirect) mechanism as is hypothesized by the sponsors of FD&C Red No. 3." [the memo, page 1]. The memo includes all of the tabular data that was used to generate a statistical determination.

I have undertaken a statistical re-examination of the data to assess whether or not the data suggest there is a biologically meaningful increase in carcinomas.

Our Observations

Our understanding of tumorigenesis and carcinogenesis has changed significantly over the years. Today, it is well-established that thyroid adenomas rarely become carcinomas, especially thyroid follicular adenomas (5% of adenomas are reported to be cancers according to StatPearls). Thus, it is wholly inappropriate to sum ademonas and carcinomas, under the assumption that adenomas will become carcinomas. Our understanding of thryoid cancers today simply does not support the notion that one should sum adenomas and carcinomas.

Sprague-Dawley rats have a relatively high background rate of follicular cell adenomas and carcinomas, according to historical control data from LabCorp published in the journal Toxicologic Pathology.

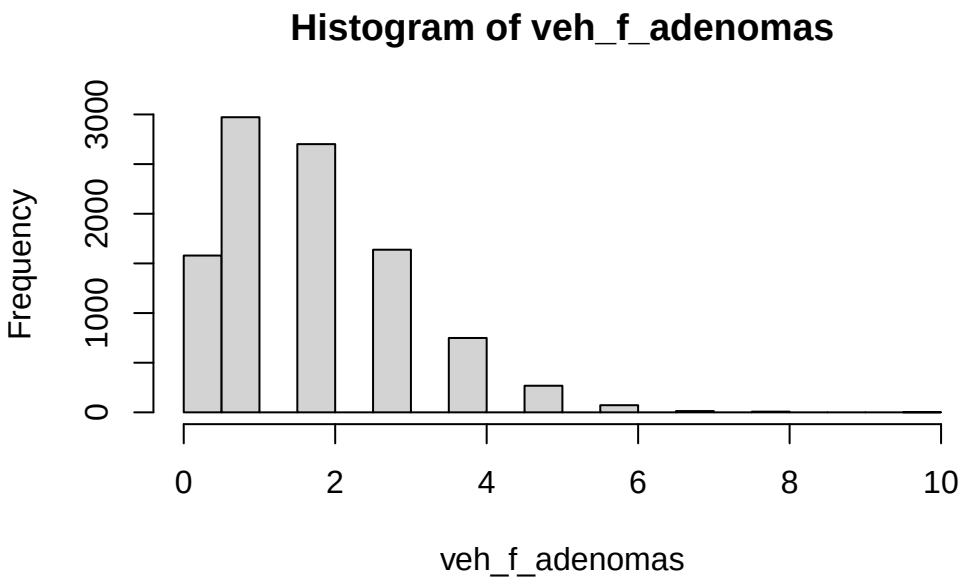
Neoplasm	Males (% of 1,801 animals)	Females (% of 1,803 animals)
Thyroid Follicular Cell Adenoma	2.8%	0.7%
Thyroid Follicular Cell Carcinoma	0.9%	0.7%

We know that the data analyzed by the US FDA in the memo suffers from sampling bias. Specifically, the sample sizes are simply too small to adequately represent the population they were drawn from. As a result, we know that we cannot draw valid inferences. Furthermore, we can see in Table 1 that there is a lack of an expected dose-response relationship. Specifically, there is a departure from a monotonic increase in adenomas and carcinomas as a function of increasing dose. Consider that the mid-dose carcinoma rate goes back down to the vehicle rate. This suggests that that what we are seeing at the low and high doses is noise hovering around the vehicle rate. That is the definition of noise.

When we have small sample sizes, we increase the likelihood of finding noise statistically significant. This is called a false positive result. We also call it "chasing noise".

We also see a very unusual lack of follicular cell adenomas in the vehicle control groups across the studies reported in the memo . Given an expected rate of 2.8% based on the LabCorp data, we would expect the vehicles to show at least 1 or 2 follicular cell adenomas on average – but that's on average. What is the probability we would have 0 follicular cell adenomas in our study? We will assume we have 65 animals in the group.

```
set.seed(530727)
veh_f_adenomas <- rbinom(10000, 65, 0.028)
hist(veh_f_adenomas)
```



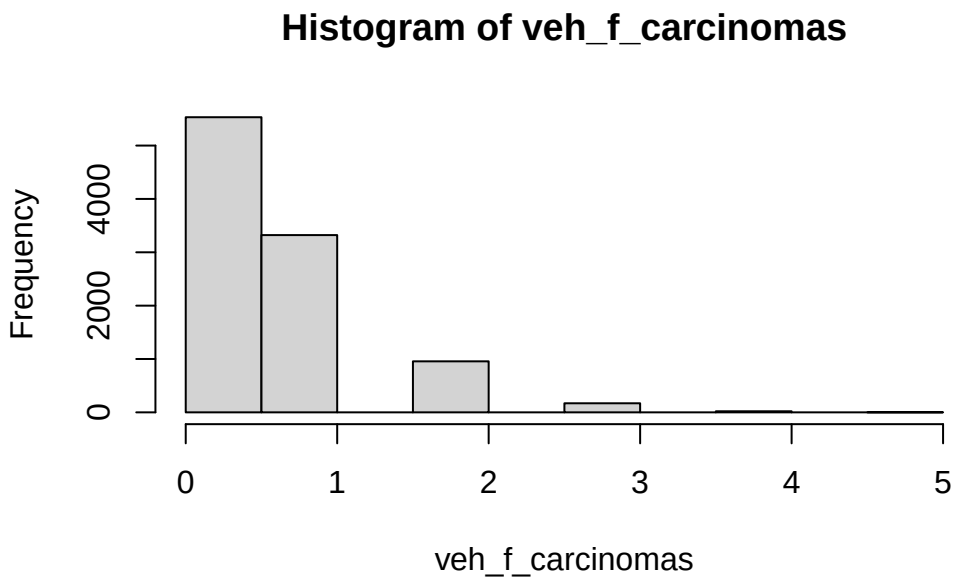

```
length(which(veh_f_adenomas == 0))/length(veh_f_adenomas)
```

```
[1] 0.1579
```

There's a 16% chance that we would have 0 follicular cell adenomas in our study. Since each vehicle control is likely independent, one would expect each to have a 16% chance of having 0 follicular cell adenomas.

We can also ask the question of what is the likelihood we would see 0 or 1 follicular carcinoma given a background rate of 0.9%? We will assume we have 65 animals in the group.

```
veh_f_carcinomas <- rbinom(10000, 65, 0.009)
hist(veh_f_carcinomas)
```



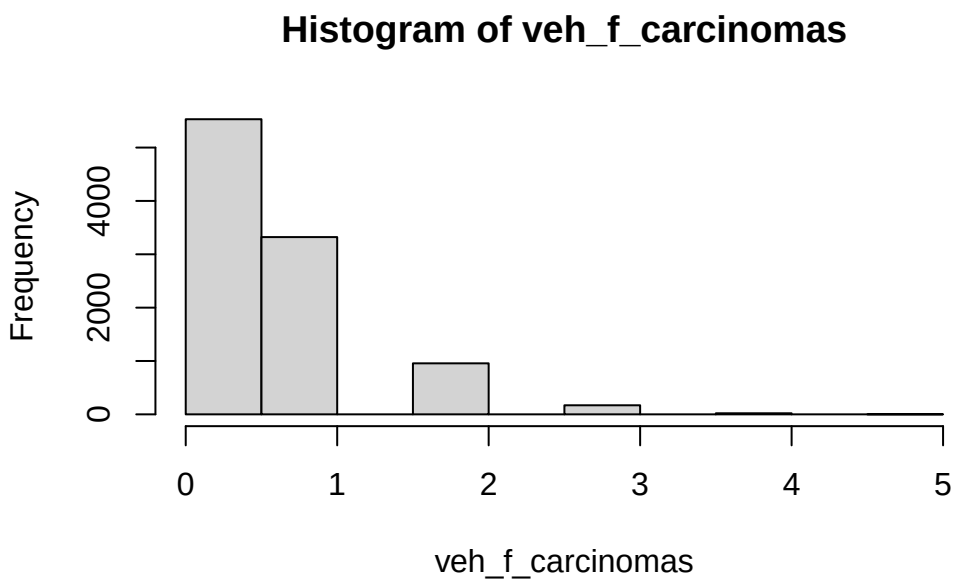

```
length(which(veh_f_carcinomas == 0))/length(veh_f_carcinomas)
```

```
[1] 0.5531
```

We can see that there's a 55% probability of having 0 follicular carcinomas in each control group.

What is the probability we would see 3/57 follicular carcinomas in the high group if it was no different from the baseline?

```
t_high_f_carcinomas <- rbinom(10000, 57, 0.009)
hist(veh_f_carcinomas)
```




```
length(which(t_high_f_carcinomas == 3))/length(t_high_f_carcinomas)
```

```
[1] 0.0161
```

So, there's a 2% chance of seeing 3/57 follicular carcinomas. That would be consistent with noise, especially after correcting for all of the multiple testing that US FDA is doing. It is important to note that the US FDA did not perform any type of multiple testing correction following all of the statistical tests they ran. With each additional test, they are increasing the false positive rate. It is a statistical analysis best practice to control the Family-Wise Error Rate when performing multiple statistical tests (i.e., they need to control for multiple tests in order to maintain a 5% false positive rate).

At issue here is that the US FDA is analyzing each study in isolation, and calculating to see if there is a significant increase in adenomas and carcinomas. We have already addressed the fact that adding adenomas and carcinomas is problematic not only from a biological standpoint (i.e., the overwhelming majority (at least 95%) of adenomas do not become carcinomas), but also from a legal standpoint – the Delaney Clause specifically states "cancer" not "adenomas" and not "benign tumors".

The 9th Circuit Court of Appeals made clear that Congress was very clear and specific in its construction of the Delaney Clause. Petitioners point this out as well. However, what petitioners failed to articulate is that because of the 9th Circruit Court of Appeals decision,

the use of adenoma data is inappropriate. Congress was specific in its construction fo the Delaney Clause when it stated "cancer". Congress did not say, "adenoma" or "benign tumor". Congress was clear that only if a chemical causes "cancer" in animals or humans should it be impacted by the Delaney Clause. And it is clear that at the time Congress passed the Delaney Clause there was an understanding that cancer was separate from benign tumors.

Thus, when US FDA took the sum of adenomas and carcinomas, it was violating the specific direction given by Congress. This strict interpretation of Congress' text and intent is further reinforced by the more recent Supreme Court decision in Loper Bright. Thus, in light of the 9th Circuit decision and in light of Loper Bright, it is abundantly clear that one is not to utilize adenoma data in any way to assess if a chemical is causing cancer in animals or humans for the purposes of the Delaney Clause.

And as already mentioned, we have biological reasons to also exclude adenomas from the calculation to assess if a chemical causes cancers in animals.

FDA takes an antiquated approach (by today's standards) to analyze the cancer data. And at that time, what they were doing was mostly state of the art. They analyzed each study individually and calculated a p-value. However, as the American Statistical Association has repeatedly cautioned, a p-value is not a bright line, nor is it an indicator of biological meaningfulness. Thus, relying strictly on a p-value for decision making is highly inappropriate. In addition, Ronald Fisher, the father of the p-value, has stated in numerous texts that all a significant p-value means is that additional testing is warranted. The reason for this is simple – 1) the result must be reproduced to ensure the result is actually correct and not a false positive, and 2) Fisher and other statisticians were aware of the concept of "regression to the mean", which states that extraordinary results, upon re-testing, will tend to move towards the mean upon further testing, and that over time with repeated testing, the most likely value (the mean) will emerge more clearly.

A more robust and modern way to analyze this data is to take a more Bayesian approach [this exact approach was available at the time the US FDA was doing its calculations; however, Bayesian approaches were not mainstream at the time]. In this case, we can use a beta-binomial analysis to more directly calculate what the likely rate of occurence is given the data that we have. This approach can combine the data from all of the studies together. The advantage of this approach is that we can get a better sense of the background/historical cancer rate in the vehicle control animals. In fact, you will see that the uncertainty in the background/historical cancer rate in the vehicle control animals has very little uncertainty compared to the data for the 4% FD&C Red No. 3 groups.

The following are the steps we will take in the analysis:

Step 1: Build a vehicle data model

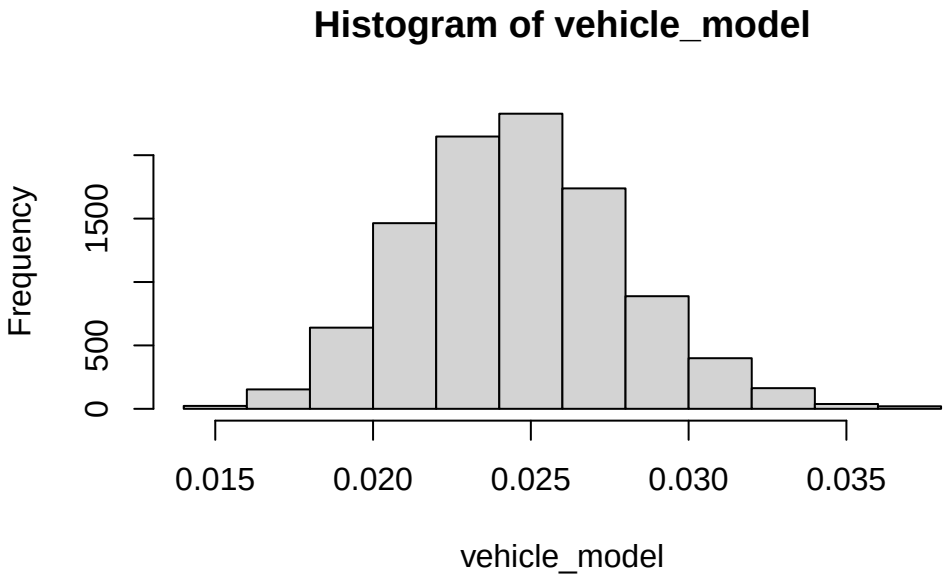
Step 2: Build a data model for the 4% group

Step 3: Determine if the models are likely different or not based on the 95% Highest Density Interval, and a region of practical equivalence of 1/58 (0.017) – since that is the smallest difference we can observe.

Analysis

The first step is to build the vehicle model:

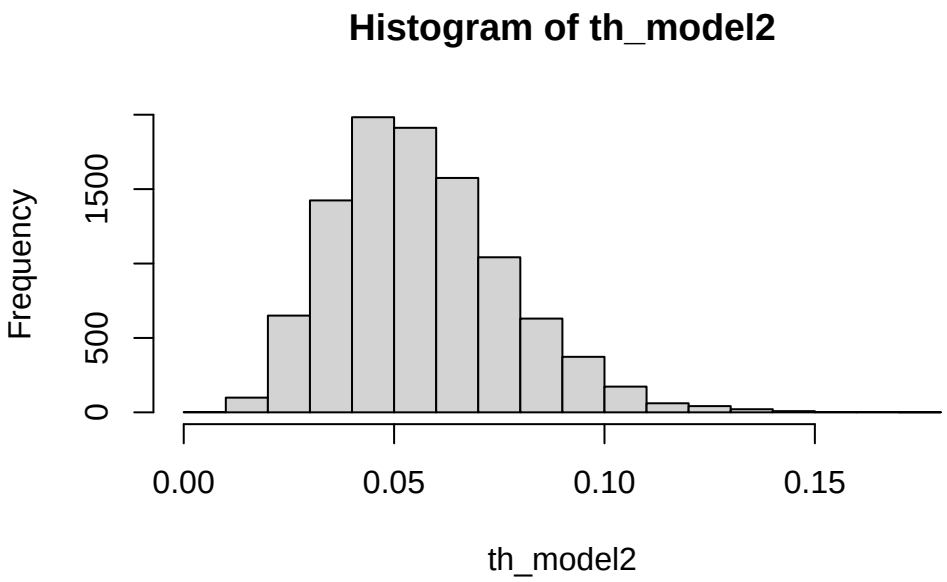
```
vehicle_model <- rbeta(10000, (1+0+1+0+50+1)-1, (59+66+65+61+1794+68-1-(1+0+0+0+1+0+50+1))
hist(vehicle_model)
```



Next, we will build a follicular carcinoma model for the 4% group: we will be using data from Tables 5 and 6 (Table 2 is female data only so it will not be used, as the FDA states that this is a male rat issue only).

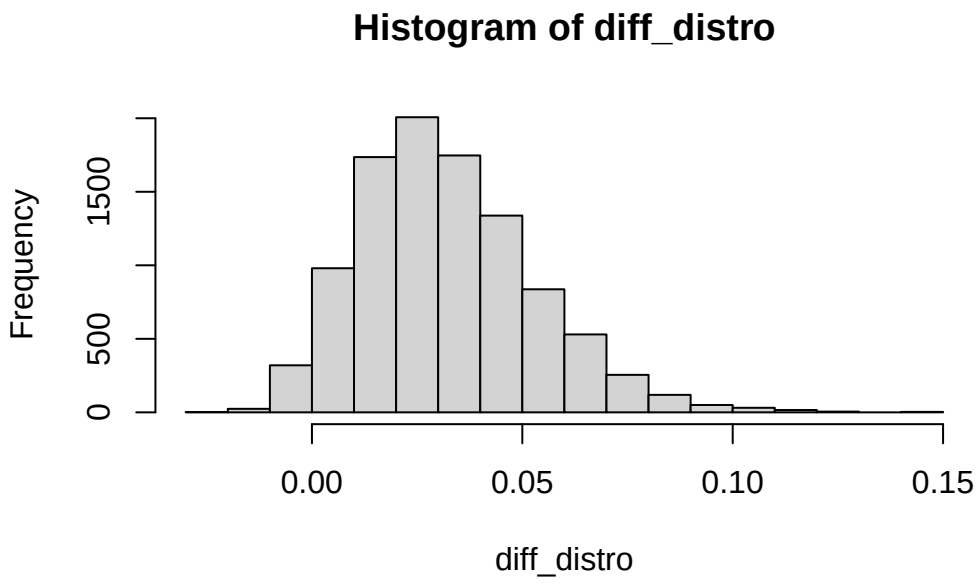
Then the 4% model would look like this:

```
th_model2 <- rbeta(10000, (5+3)-1, (68+58-(5+3)-1))
hist(th_model2)
```



The difference distribution would look like this:

```
diff_distro <- th_model2 - vehicle_model
hist(diff_distro)
```



And that would leave a 95% Highest Density Interval of:

```
library(HDInterval)
hdi(diff_distro)
```

```
       lower        upper
-0.006736218  0.071494675
attr(,"credMass")
[1] 0.95
```

This still includes the value of 0.

```
median(vehicle_model)
```

```
[1] 0.02447365
```

```
median(th_model2)
```

```
[1] 0.05428105
```

```r
median(diff_distro)
```

```
[1] 0.02972931
```

So now the median of the 4% group's distribution is 5.4% instead of 3.5%. The value of 0 is still within the 95% HDI of the difference distribution – meaning 0 difference is a plausible effect size.

Assumptions

The primary assumption of this analysis is that all of the studies are exchangeable. This is a valid assumption as if it were not, the US FDA would not be able to compare the studies. Furthermore, the historical background data are exchangeable with these studies as it was conducted by a mainstream Contract Research Organization that provides similar types of analyses using the same animals from the same vendor. If there were variables that impacted thyroid tumor development significantly beyond the test chemical that were not shared across the studies, then no study would be reliable.

We are making the assumption that the studies themselves are reliable to perform the analysis. We know this is not the case as the sample sizes are too small to invoke the Central Limit Theorem. We also know that small sample sizes do not replicate the population they are drawn from. We assume that we can do our best to estimate the population distribution based on these small sample sizes; however, we also know that it's simply not possible given the few studies we have available to us.

Conclusions

When taking all of the studies together, considering them, and building statistical models from them, we can see that the data for the 4% group is not meaningfully different from the vehicle controls, as the vehicle control distribution is a subset of the 4% group's distribution. This is evidenced through the inclusion of 0% difference in the difference distribution and the 95% Highest Density Interval of the difference distribution.

The US FDA should have ruled that the data currently available on FD&C Red No. 3, relied upon by the petitioners is unreliable due to sampling bias. In fact, these results demonstrate the fact, when looking at the large amount of uncertainty in the 4% group distribution, that the data are unreliable. This is a stance that the US FDA has taken in the past with other Delaney Clause decisions, and it is the stance US FDA should have taken on this petition (e.g., US FDA approved N,N,N',N',N'',N''- hexakis(methoxymethyl)-1,3,5-triazine-2,4,6-triamine polymer with stearyl alcohol, -octadecenyl--hydroxypoly(oxy-1,2- ethanediyl), and alkyl (C20+)

alcohols as a component of paper and paperboard in contact with aqueous foods despite the presence of certain impurities, in part based on the fact that the cancer data suggesting one impurity causes cancer was unreliable). [https://www.govinfo.gov/content/pkg/FR-1994-09-21/html/94-23274.htm].

Thus, we have illustrated 2 different ways of coming at the same conclusion –the US FDA should have rejected the remedy sought by petitioners for 1) relying upon unreliable data, 2) relying upon data that showed there was too much uncertainty to draw a conclusion (thus, the data were unreliable), and 3) when looking at all of the data together, it is clear that there is no evidence that the 4% group caused cancers at a higher rate than the vehicle controls (i.e., FD&C Red No. 3 does not cause cancer at a rate that is biologically meaningfully different from the background rate).

Furthermore, the US FDA erred when it combined adenomas and carcinomas. There is no biological basis for continuing that practice, as over 95% of adenomas do not become carcinomas. In addition, the Delaney Clause is quite clear, and both the 9th Circuit and the Supreme Court have stated that a strict interpretation of Congress' text and intent are required. Congress was clear that it only wanted to restrict dyes that cause "cancer". Not dyes that cause "benign tumors". Not dyes that cause "adenomas". These are all words that were well known at the time that Congress wrote and passed the amendment that included the Delaney Clause. Therefore, Congress was clear and unambiguous – only "cancer". Therefore, when US FDA combined adenomas and carcinomas to come to a conclusion, it erred and violated the clear direction given by Congress.

These are my opinions to an appropriate degree of scientific, pharmacological, and toxicological certainty.

Lyle D. Burgoon, Ph.D., Fellow ATS

President and CEO

Raptor Pharm & Tox, Ltd.

Dr. Burgoon's Background

Dr. Burgoon is an award-winning and Board Certified (Fellow, Academy of Toxicological Sciences) toxicologist and biostatistician. Dr. Burgoon has 10+ years of experience as a US Government toxicologist employed by the US Environmental Protection Agency (US EPA)

and the US Army. In 2022, Dr. Burgoon was recognized for his expertise in children's environmental health when US EPA Administrator Regan appointed Dr. Burgoon to the Children's Health Protection Advisory Committee (CHPAC), with his term beginning in 2023.

During Dr. Burgoon's prior service at the US EPA, he served as a senior leader in various roles from leading teams of biostatisticians and bioinformaticians in designing and analyzing toxicology studies, to supervising toxicologists and public health scientists in the development of toxicological and dose-response assessments in support of the US EPA Integrated Risk Information System (IRIS). At the US Army, Dr. Burgoon led classified and unclassified programs and projects with a focus on the effects of chemical, biological, radiological, nuclear, and high yield explosives (CBRNE) weapons on human health, ecological and agriculture systems.

Dr. Burgoon's expertise is in pharmacology, toxicology, biostatistics, exposure modeling, causal analysis, risk assessment, especially in the area of establishing causal relationships of chemical and microbiological exposures to human effects and the development and application of mathematical modeling, machine learning solutions for computational pharmacology and toxicology. Dr. Burgoon has knowledge, training, and experience in the toxicological impacts of drugs and chemicals (including alcohol), biological agents, food safety, weapons of mass destruction, environmental stressors, and physiological responses to a variety of stressors. Dr. Burgoon also has knowledge, training, and experience in analytical chemistry associated with drug and alcohol testing. In addition, Dr. Burgoon's training, knowledge, and experience in biostatistics gives him the ability to assess scientific study designs, analyses, and overall study quality.

Dr. Burgoon is well-published with articles appearing in peer-reviewed publications and he has coauthored book chapters. Dr. Burgoon has extensive experience as a consultant, leading his own consulting practice, as well as serving as a toxicology and risk assessment consultant at the US EPA, and a toxicology, biostatistics, and risk assessment consultant to organizations across the US Department of Defense and the US Army, and serving as a study section/special emphasis panel reviewer for the US National Institutes of Health.